

# Investigating Pertussis Resurgence

Jie

Q1. With the help of the R “addin” package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

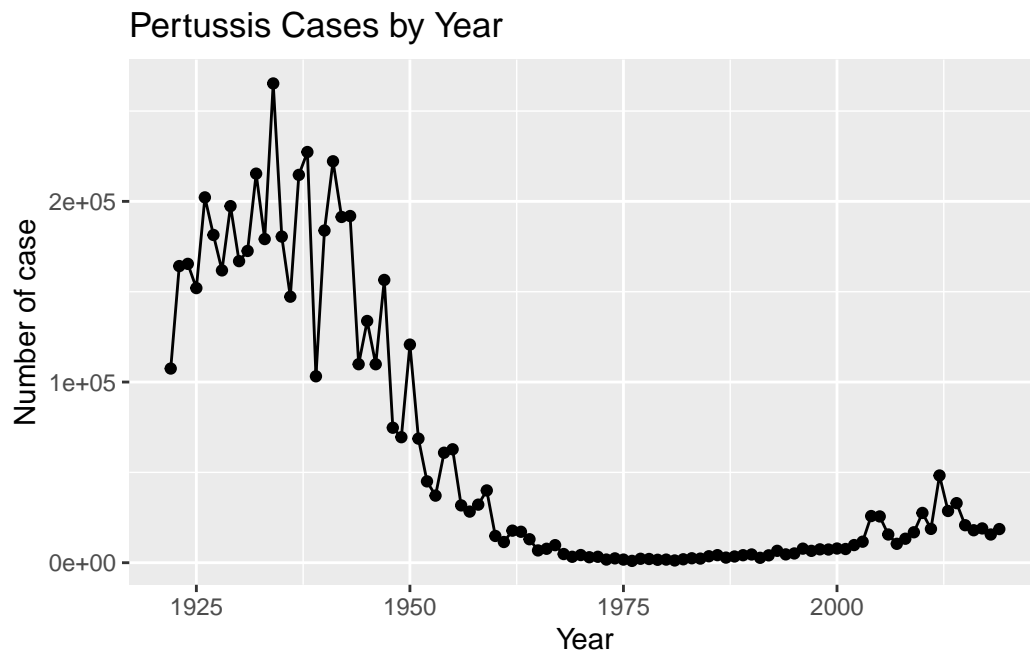
```
cdc <- data.table::data.table(  
  Year = c(1922L,1923L,1924L,1925L,  
           1926L,1927L,1928L,1929L,1930L,1931L,  
           1932L,1933L,1934L,1935L,1936L,  
           1937L,1938L,1939L,1940L,1941L,1942L,  
           1943L,1944L,1945L,1946L,1947L,  
           1948L,1949L,1950L,1951L,1952L,  
           1953L,1954L,1955L,1956L,1957L,1958L,  
           1959L,1960L,1961L,1962L,1963L,  
           1964L,1965L,1966L,1967L,1968L,1969L,  
           1970L,1971L,1972L,1973L,1974L,  
           1975L,1976L,1977L,1978L,1979L,1980L,  
           1981L,1982L,1983L,1984L,1985L,  
           1986L,1987L,1988L,1989L,1990L,  
           1991L,1992L,1993L,1994L,1995L,1996L,  
           1997L,1998L,1999L,2000L,2001L,  
           2002L,2003L,2004L,2005L,2006L,2007L,  
           2008L,2009L,2010L,2011L,2012L,  
           2013L,2014L,2015L,2016L,2017L,2018L,  
           2019L),  
  Cases = c(107473,164191,165418,152003,  
            202210,181411,161799,197371,  
            166914,172559,215343,179135,265269,  
            180518,147237,214652,227319,103188,  
            183866,222202,191383,191890,109873,  
            133792,109860,156517,74715,69479,  
            120718,68687,45030,37129,60886,  
            62786,31732,28295,32148,40005,
```

```
14809,11468,17749,17135,13005,6799,
7717,9718,4810,3285,4249,3036,
3287,1759,2402,1738,1010,2177,2063,
1623,1730,1248,1895,2463,2276,
3589,4195,2823,3450,4157,4570,
2719,4083,6586,4617,5137,7796,6564,
7405,7298,7867,7580,9771,11647,
25827,25616,15632,10454,13278,
16858,27550,18719,48277,28639,32971,
20762,17972,18975,15609,18617)
```

```
)
```

```
library(ggplot2)
```

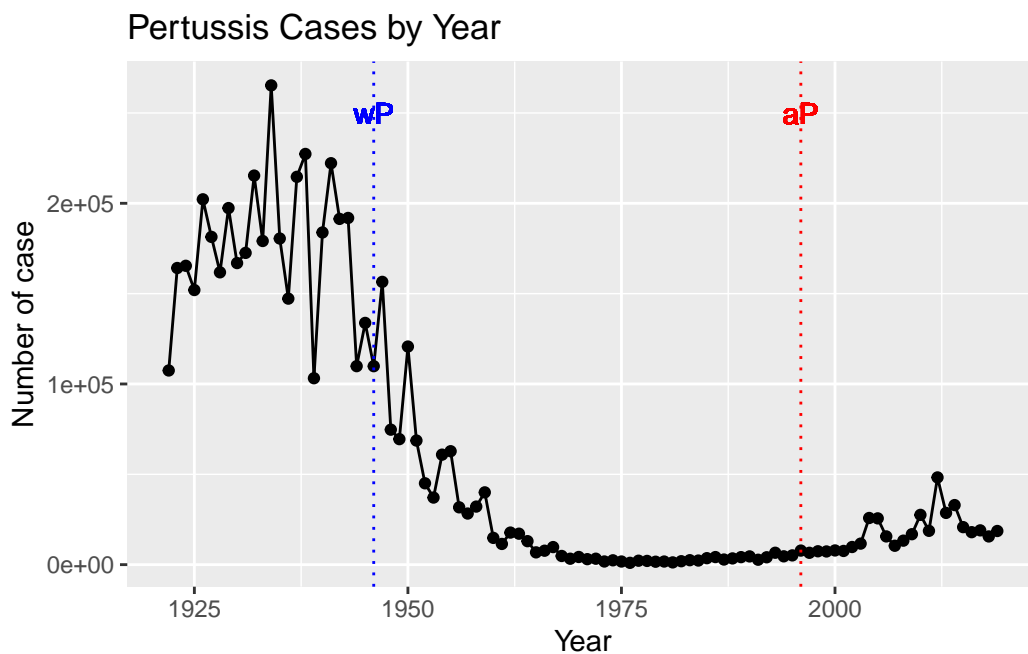
```
ggplot(cdc) +
  aes(x=Year, y=Cases) +
  geom_point() +
  geom_line()+
  labs(title="Pertussis Cases by Year",x="Year",y="Number of case")
```



Q2. Using the ggplot `geom_vline()` function add lines to your previous plot for the 1946

introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```
ggplot(cdc) +  
  aes(x=Year, y=Cases) +  
  geom_point() +  
  geom_line()+  
  geom_vline(xintercept=1946,color="blue",linetype="dotted")+  
  geom_vline(xintercept=1996,color="red",linetype="dotted")+  
  geom_text(x=1946,y=250000,label="wP",color="blue")+  
  geom_text(x=1996,y=250000,label="aP",color="red")+  
  labs(title="Pertussis Cases by Year",x="Year",y="Number of case")
```



Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

It is clear from the CDC data that pertussis cases are once again increasing. For example, we can see that in 2012 the CDC reported 48,277 cases of pertussis in the United States. This is the largest number of cases reported since 1955, when 62,786 cases were reported. The pertussis field has several hypotheses for the resurgence of pertussis including (in no particular order): 1) more sensitive PCR-based testing, 2) vaccination hesitancy 3) bacterial evolution (escape from vaccine immunity), 4) waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.

```
library(jsonlite)
```

```
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
```

```
head(subject, 3)
```

```

  subject_id infancy_vac biological_sex ethnicity race
1          1          wP      Female Not Hispanic or Latino White
2          2          wP      Female Not Hispanic or Latino White
3          3          wP      Female      Unknown White
  year_of_birth date_of_boost   dataset
1 1986-01-01    2016-09-12 2020_dataset
2 1968-01-01    2019-01-28 2020_dataset
3 1983-01-01    2016-10-10 2020_dataset

```

Q4. How many aP and wP infancy vaccinated subjects are in the dataset?

```
table(subject$infancy_vac)
```

```

aP wP
47 49

```

Q5. How many Male and Female subjects/patients are in the dataset?

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

```
table(subject$biological_sex, subject$race)
```

```

      American Indian/Alaska Native Asian Black or African American
Female                0      18                2
Male                  1       9                0

      More Than One Race Native Hawaiian or Other Pacific Islander
Female                8                1
Male                  2                1

```

	Unknown	or Not Reported	White
Female	10	27	
Male	4	13	

```
library(lubridate)
```

Attaching package: 'lubridate'

The following objects are masked from 'package:base':

date, intersect, setdiff, union

```
today()
```

```
[1] "2022-12-01"
```

```
today() - ymd("2000-01-01")
```

Time difference of 8370 days

```
time_length( today() - ymd("2000-01-01"), "years")
```

```
[1] 22.91581
```

Q7. Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different? Hint Using ymd() function on the appropriate column of our subject data frame together with today's date we can calculate the age in days for subjects. We can optionally store this back into our subject data frame for later use:

## Use today's date to calculate age in days

```
subject$age <- today() - ymd(subject$year_of_birth)
```

Next we can use dplyr's filter() function to limit ourselves to a particular subset of subjects to examine the 6 number summary of their age in years:

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
ap <- subject %>% filter(infancy_vac == "aP")
round( summary( time_length( ap$age, "years" ) ) )
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
23	25	26	25	26	27

```
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28	32	35	36	40	55

Q8. Determine the age of all individuals at time of boost?

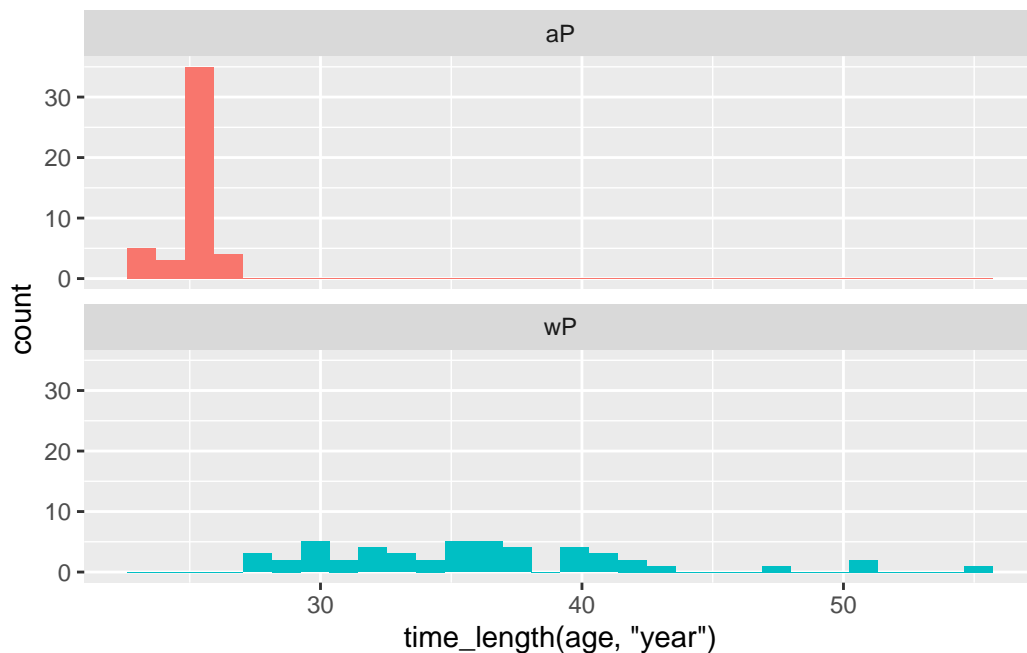
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

```
ggplot(subject) +  
  aes(time_length(age, "year"),  
       fill=as.factor(infancy_vac)) +  
  geom_histogram(show.legend=FALSE) +  
  facet_wrap(vars(infancy_vac), nrow=2)
```

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)  
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

Q9. Complete the code to join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details:

```
meta <- full_join(specimen, subject)
```

```
Joining, by = "subject_id"
```

```
dim(meta)
```

```
[1] 729 14
```

```
head(meta)
```

```
specimen_id subject_id actual_day_relative_to_boost
1           1           1                        -3
2           2           1                       736
3           3           1                        1
4           4           1                        3
5           5           1                        7
6           6           1                       11
planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                           0         Blood     1          wP        Female
2                          736         Blood    10          wP        Female
3                           1         Blood     2          wP        Female
4                           3         Blood     3          wP        Female
5                           7         Blood     4          wP        Female
6                          14         Blood     5          wP        Female
ethnicity race year_of_birth date_of_boost dataset
1 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
2 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
3 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
4 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
5 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
6 Not Hispanic or Latino White 1986-01-01 2016-09-12 2020_dataset
age
1 13483 days
2 13483 days
3 13483 days
4 13483 days
5 13483 days
6 13483 days
```

Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.



```
abdata <- inner_join(titer, meta)
```

Joining, by = "specimen\_id"

```
dim(abdata)
```

```
[1] 32675    21
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
table(abdata$isotype)
```

```

IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 1413 6141 6141 6141 6141

```

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

```
table(abdata$visit)
```

```

 1    2    3    4    5    6    7    8
5795 4640 4640 4640 4640 4320 3920   80

```

```

ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)

```

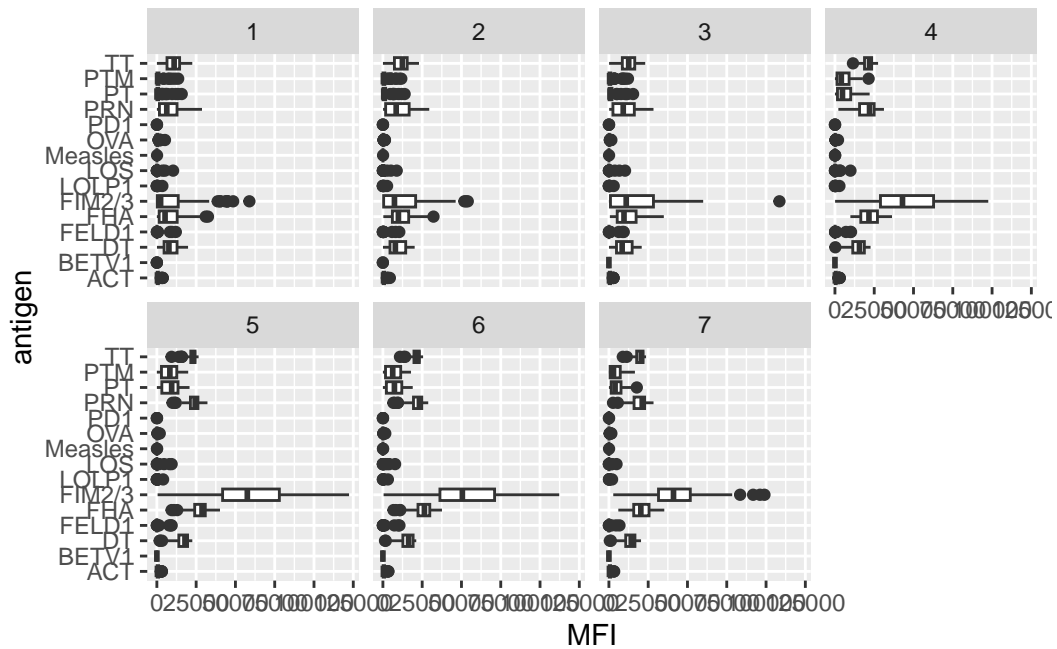
	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG1	TRUE	ACT	274.355068	0.6928058
2	1	IgG1	TRUE	LOS	10.974026	2.1645083
3	1	IgG1	TRUE	FELD1	1.448796	0.8080941
4	1	IgG1	TRUE	BETV1	0.100000	1.0000000
5	1	IgG1	TRUE	LOLP1	0.100000	1.0000000
6	1	IgG1	TRUE	Measles	36.277417	1.6638332

unit lower\_limit\_of\_detection subject\_id actual\_day\_relative\_to\_boost

1	IU/ML	3.848750	1		-3
2	IU/ML	4.357917	1		-3
3	IU/ML	2.699944	1		-3
4	IU/ML	1.734784	1		-3
5	IU/ML	2.550606	1		-3
6	IU/ML	4.438966	1		-3
	planned_day_relative_to_boost	specimen_type	visit	infancy_vac	biological_sex
1	0	Blood	1	wP	Female
2	0	Blood	1	wP	Female
3	0	Blood	1	wP	Female
4	0	Blood	1	wP	Female
5	0	Blood	1	wP	Female
6	0	Blood	1	wP	Female
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
5	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
6	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
	age				
1	13483 days				
2	13483 days				
3	13483 days				
4	13483 days				
5	13483 days				
6	13483 days				

Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

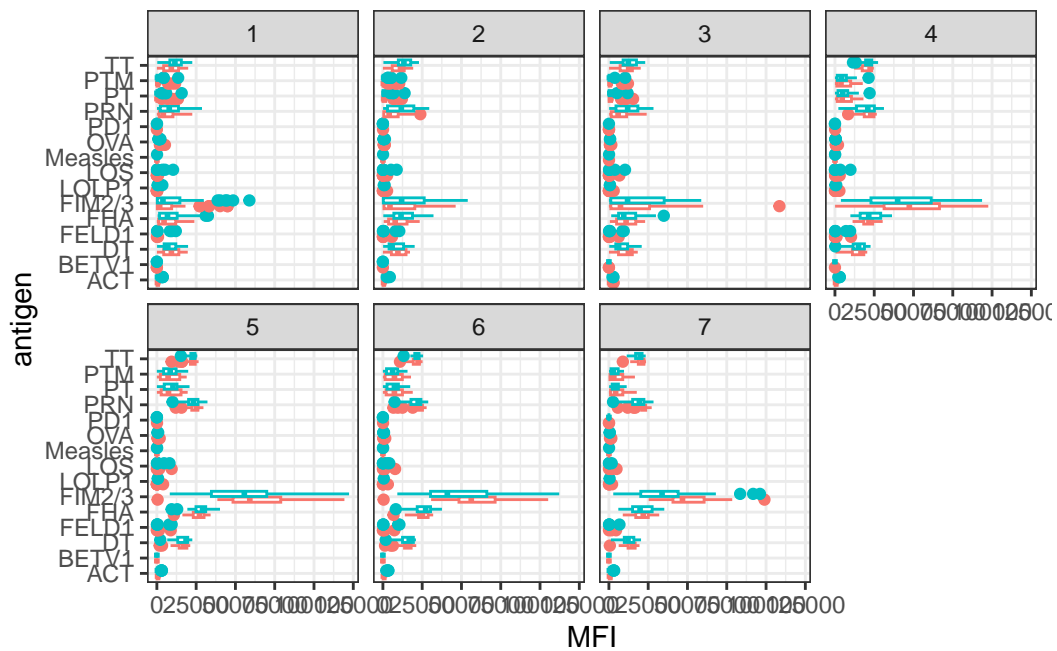
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



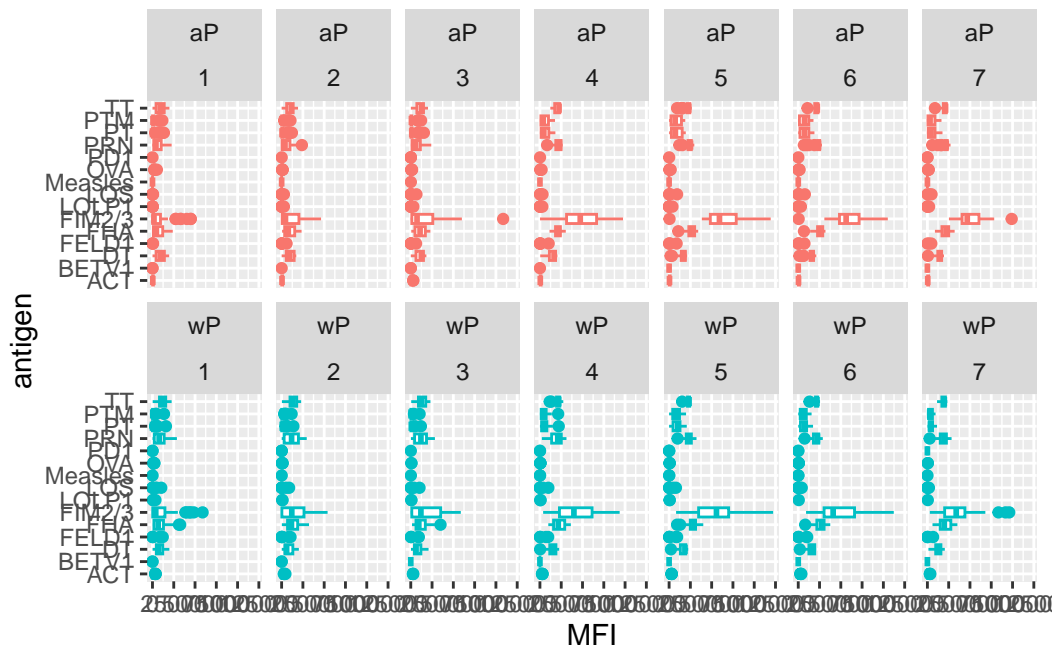
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

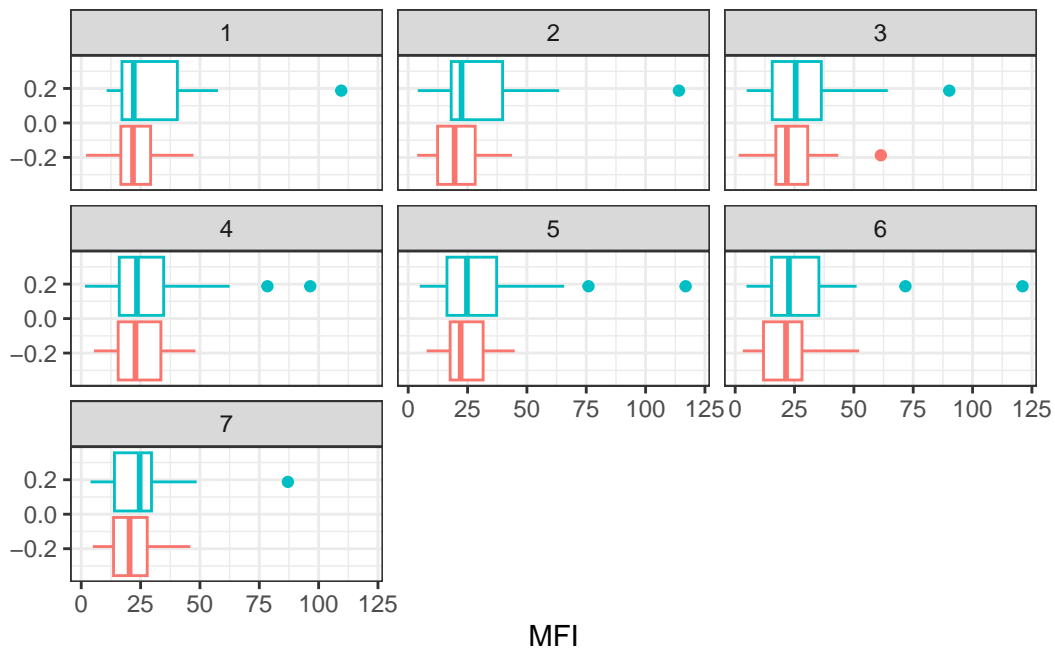


```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

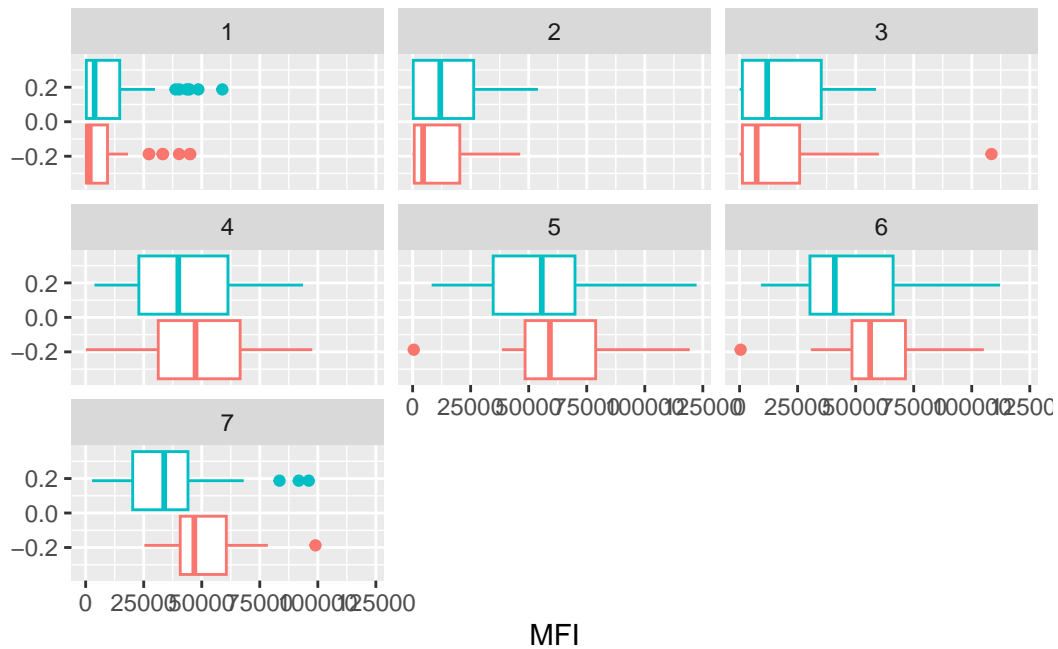


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a “control” antigen (“Measles”, that is not in our vaccines) and a clear antigen of interest (“FIM2/3”, extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit))
```



```

aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()

```

NULL

Q16. What do you notice about these two antigens time courses and the FIM2/3 data in particular? Hint FIM2/3 levels clearly rise over time and far exceed those of Measles. They also appear to peak at visit 5 and then decline. This trend appears similar for for wP and aP subjects.

Q17. Do you see any clear difference in aP vs. wP responses?

```

url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSOG00000211896."

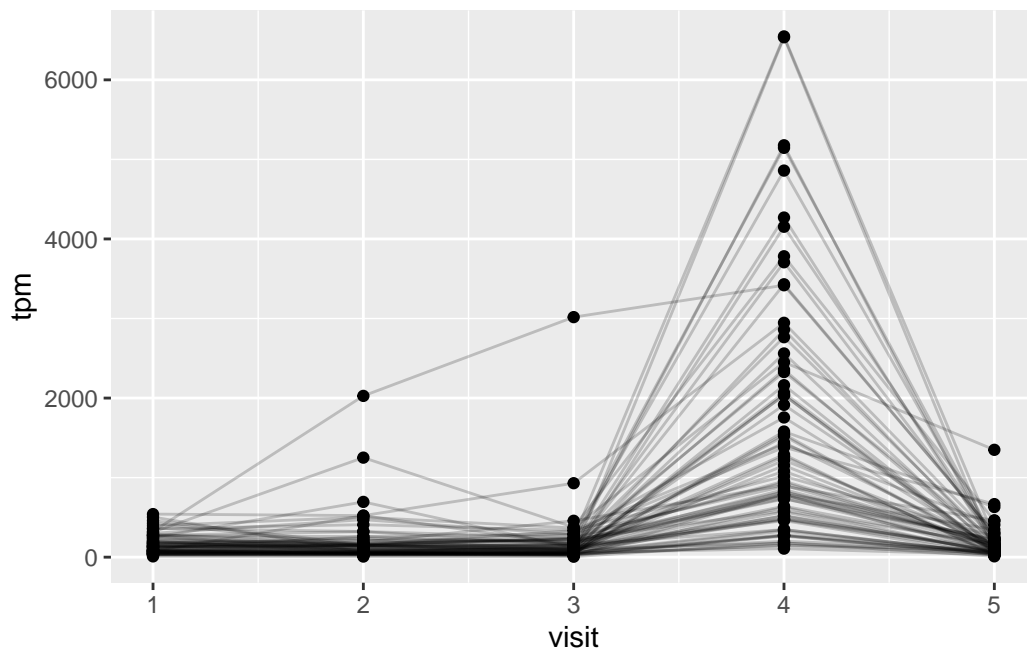
rna <- read_json(url, simplifyVector = TRUE)

ssrna <- inner_join(rna, meta)

```

Joining, by = "specimen\_id"

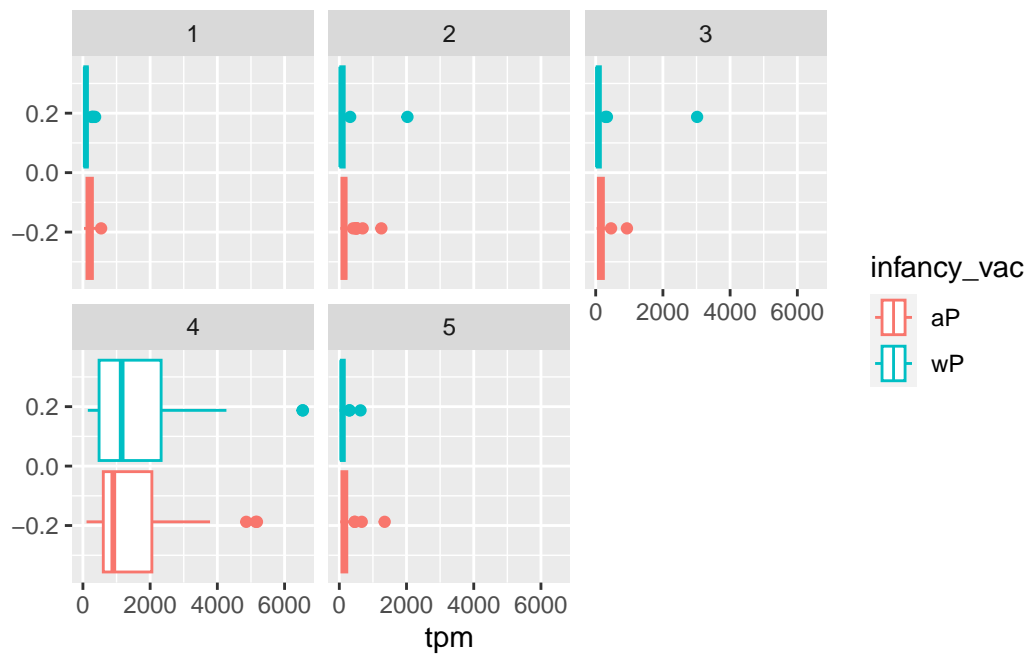
```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)? When the visit is 4

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not? Ydx  
I think tehy match -

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```



